Sparse Autoencoders are Capable LLM Jailbreak Mitigators

Yannick Assogba, Jacopo Cortellazzi, Javier Abad, Pau Rodriguez, Xavier Suau, Arno Blaas
Mechanistic Interpretability Workshop ICML 2026 · Apple, EPFL

Summary

Prompt based jailbreak attacks remain a persistent threat to large language model safety. We propose **Context-Conditioned Delta Steering (CC-Delta)**, an SAE-based defense that identifies jailbreak-relevant sparse features by comparing token-level representations of the same harmful request *with and without* jailbreak context.

We compare the safety gain vs utility loss (as measured by IFEval, MMLU and Fluency) and find that **CC-Delta** outperforms dense mean-shift steering on all models in our test suite, especially against out-of-distribution and adaptive attacks.

CC-Delta

1. Token Matched Diffs

$$\underbrace{\frac{1}{T(i)}\sum_{t=1}^{T(i)}\mathbf{z}_t^{(i)}}_{\text{harmful request tokens}} - \underbrace{\frac{1}{T(i)}\sum_{t=1}^{T(i)}\bar{\mathbf{z}}_t^{(i)}}_{\text{matched tokens in jailbreak}}$$

2. Feature Filtering & Ranking

Select significant feature diffs via Wilcoxon Signed Rank Test. Rank by standardized feature shift.

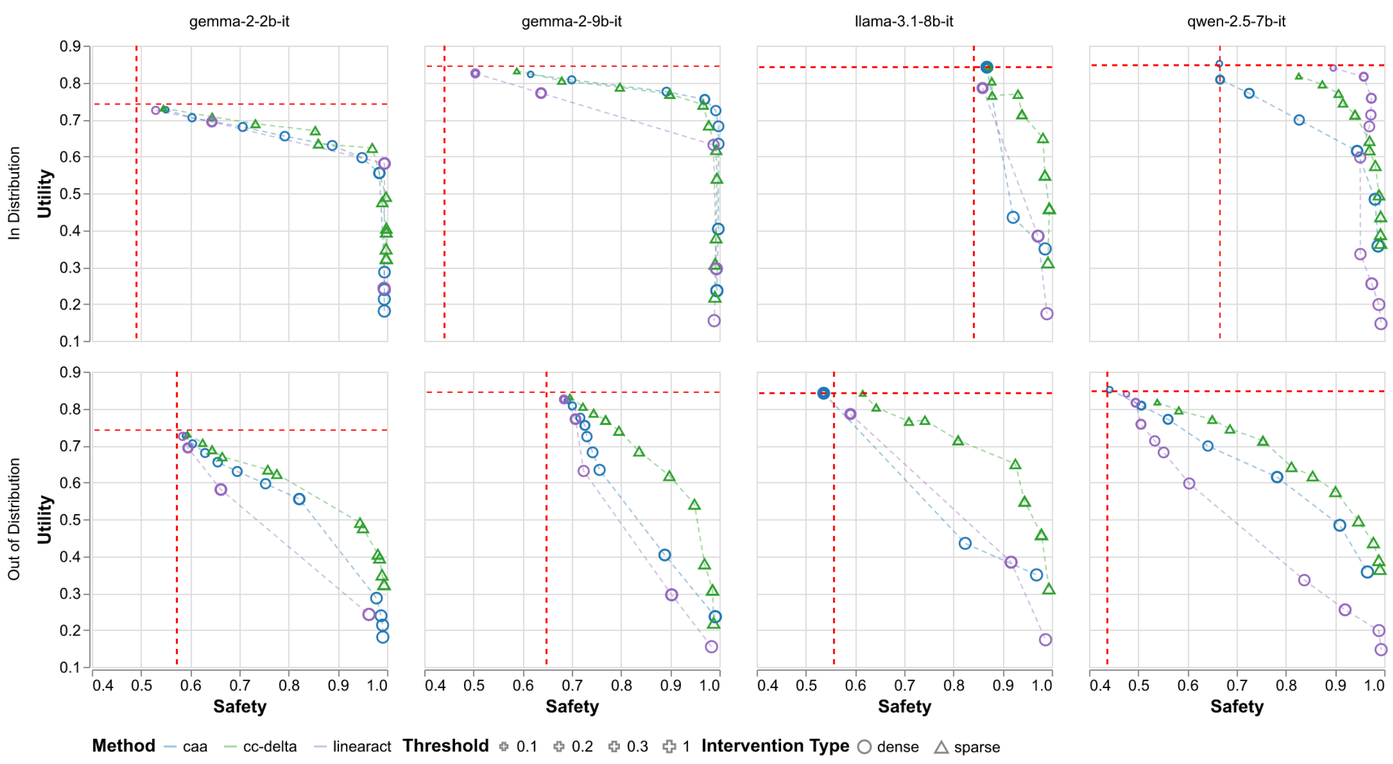
3. SAE Mean Shift Steering

Select top-N features and apply a scaled add to SAE encoded representation

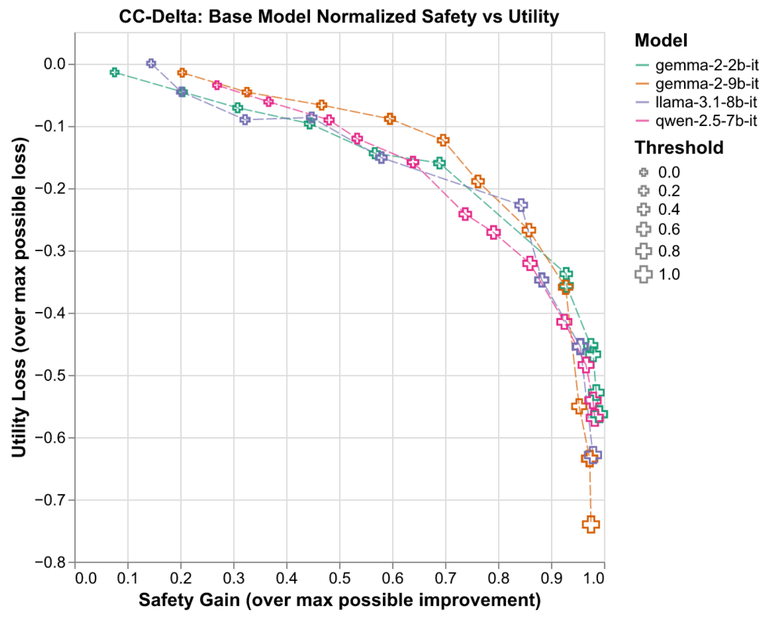
$$\textbf{Encode + Shift:}\quad \mathbf{z}' = \mathrm{SAE}(\mathbf{h}) + \alpha(\mathbf{m}\odot\boldsymbol{\Delta})$$
$$\textbf{Decode:}\quad \tilde{\mathbf{h}} = \mathrm{SAE}^{-1}(\mathbf{z}') + \mathbf{e}$$
$$\mathbf{e} = \mathbf{h} - \mathrm{SAE}^{-1}(\mathrm{SAE}(\mathbf{h})),$$

Results

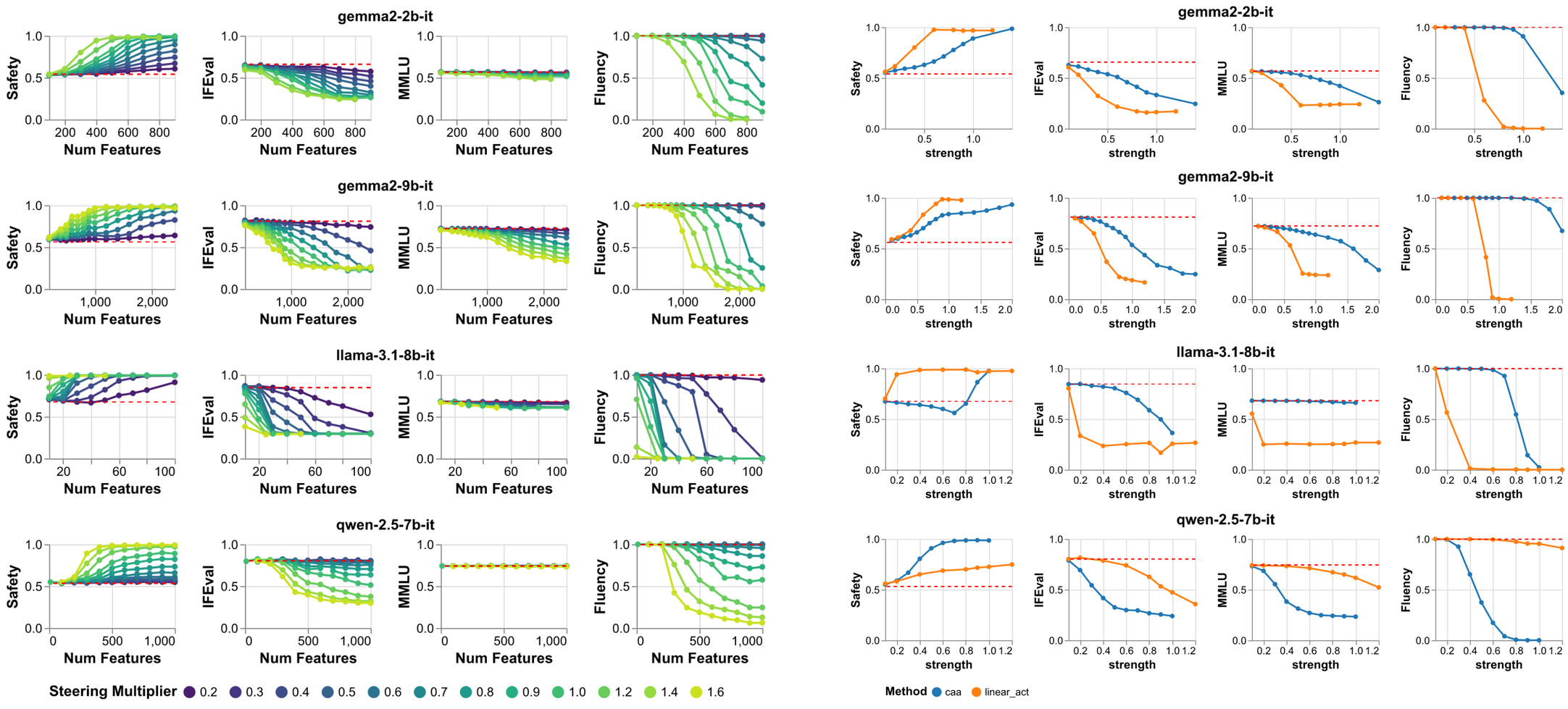
CC-Delta improves the safety–utility tradeoff over dense steering baselines



CC-Delta performs consistently across models



CC-Delta provides fine-grained control over the safety–utility tradeoff via its two main inference-time parameters.
Effective interventions often require steering *many SAE features* simultaneously.



Additional findings

- CC-Delta's gains come from context-conditioned feature selection, rather than SAE-space mean-shift steering alone.
- Features selected by CC-Delta from non-adaptive jailbreak prompts generalize to PAIR, a strong adaptive attack.
- Aggregate SAE metrics such as sparsity (L0) and reconstruction quality (FVU) did not reliably predict steering performance.